Megabase Sequencing of Mycobacterial Genomes.

Douglas Smith, P. Richterich, R. Lundstrom, P. Rice, H. Safer, M. Rubenfield, H.-M. Lee, C. Butler, Q. Xu, K. Gunderson, M. Chung, J. Maher, J. Imrich, C. Tulig, and Jen-i Mao. Collaborative Research, Inc., 1365 Main Street, Waltham, MA 02154 USA

The aim of this project is to develop technological improvements for large-scale genome sequencing using computer-assisted multiplex sequencing techniques. Our current focus is on the genomes of two important bacterial pathogens: *Mycobacterium leprae* and *Mycobacterium tuberculosis*. Over the past two years, we have generated, entered into our computer system, proofread, and assembled* over 6 Mb of raw sequencing data on 18 *M. leprae* cosmids and one *M. tuberculosis* cosmid. These sequences, totaling over 700 kb, have been analyzed for known genes and open reading frames of high coding potential. Sequence data on another set of 17 cosmids is currently being generated. So far the effort has revealed many genes of interest from the perspectives of drug resistance and therapeutics development. Significant discoveries include several mycobacterial antigens, multi-drug resistance gene homologues, methylases, sugar-transferases and a family of genes related to fatty acid synthases that may be involved in cell wall biosynthesis. In addition, we have identified a homologue of a virulence gene in Salmonella which also corresponds to a major mycobacterial antigen. Our technology development goals include substantial programmed increases in sequencing speed over the next few years. Current efforts encompass the areas of sequencing chemistry, gel technology, and informatics. A brief summary of current work in these areas will be presented.

*using the programs REPLICA™ and GTAC™ licensed from Harvard Medical School. REPLICA and GTAC were developed by Leon Mintz, Gary Gryan and George Church with support from Howard Hughes Medical Institute.